一种基于 Hash 函数抽样的数据流聚类算法①

A DataStream Clustering Algorithm Based on Hash Sampling

张 驹 黄汉永 肖 杰 (中南大学 信息科学与工程学院 湖南 长沙 410083)

摘 要: 近几年来由于数据流应用的大量涌现，基于数据流的数据挖掘算法已成为重要的研究课题，而现有的数据流聚类算法 Clustream 算法存在效率低，对大数据集适应性差等严重不足，本文提出了一种基于 Hash 函数抽样的数据流聚类算法。算法采用等时间跨度滑动窗口的思想，对每个窗口内的数据首先用 Hash 函数进行抽样，抽样后的数据先保存在存储池中，然后分析样本数据的变化情况，再利用 PAM 算法得到最终的聚类结果。从对真实数据集的实验结果上来看，算法具有良好的可行性和有效性，且在大规模数据处理的情况下，效率远高于 Clustream 算法。

关键词: 数据流聚类 抽样 Hash 函数 滑动窗口 存储池

1 引言

数据流管理与分析是数据挖掘研究领域的热点问题。数据流聚类作为知识发现的重要手段得到了深入研究，与常规聚类一样，数据流聚类也是将指定数据集划分为若干个不相交部分的过程，但是巨大的数据量与在线的处理需求使得常规的聚类算法难以在数据流上直接应用，所以提出针对数据流的聚类算法是有用且必要的。

目前已经有了一些专门针对数据流设计的聚类算法，如 CluStream 算法[1]，HPStream 算法[2]等。但是，这些算法都在数据处理能力方面有些缺点，并不能得到很好的聚类结果。CluStream 算法是一种很经典的数据流聚类算法。该算法是采用一个滑动窗口模型获取当前数据流的特征[3]：根据 K-means 算法的思想，利用每一个到达滑动窗口的数据都与最近子聚类中心的距离以及一个预定的距离阀值来判断数据是否属于该子聚类。这种处理方式简单且直接，但是最大的不足之处在于，子聚类的半径随着数据的流入而不断增大，由于没有在线淘汰“老数据”，最后导致数据量越来越大，增加了处理代价。

HSCS 算法同样采用滑动窗口的思想。当一批数据流入到滑动窗口中后，使用 Hash 函数对这些数据随机抽样；然后，将抽样后的数据转化为静态数据，采用 PAM 算法进行聚类，聚类后的数据存放在存储池中。随着时间的增加，对随后流入的数据采用同样的操作。与 Clustream 算法相比，HSCS 算法大大的减少了数据处理量，节省了处理时间。同时，处理过程中所占用的内存空间也大大减少，提高了算法的可执行性。所以，HSCS 算法的提出是有意义的。

2 基本概念及问题描述

由于处理总体数据需要的计算量太大，而且由于数据噪声的存在，使得计算结果也不完全精确，所以本文提出的 HSCS 算法在随机抽样[4]的基础上，对总体数据处理得到近似的结果。

设要聚类的总体数据为：$X=(X_{ij})_{n*m}$ 。即共有 m 个变量，n 行数据。为了简化问题，本文作以下假定：假定 m 个变量中有以下几种类型：

a) 连续型，如重量和高度等，其距离计算方法一般用欧式距离。

b) 二元型，即变量只有两个取值状态，如男和女。

c) 标称型，其状态多于二个，如颜色。

2.1 各类型变量分布函数的估计

对于分布函数的估计采用简单随机抽样，设简单随机的样本数据为 s，各类型变量的分布函数的估计有以下性质：

① 收稿时间:2008-09-06

① 对于连续随机变量，其估计分布函数为近似正态分布 $N(x_{mean},s_x^2)$，分布函数为：

$$F(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}s_x}\exp[-\frac{(y-x_{mean})^2}{2s_x^2}]dy \quad (1)$$

② 对于二元变量 x，设其状态为 0，1.所抽 S 个样本中，0 状态的个数为 s_0，1 状态的个数为 s_1，令 $p=s_0/s_1$，则其估计分布函数为：

$$F(x)=\begin{cases}p & (x=0)\\ 1 & (x=1)\end{cases} \quad (2)$$

③ 对于标称变量，设状态有 $state_1$，$state_2$，…，$state_t$，分别被标记为 1，2，…，t。所抽样本中各状态出现的个数分别为 $kstate_1$，$kstate_2$，…，$kstate_t$，令 $p_i=\frac{kstate_i}{s}$（$i=1,2,\ldots,t$）。则其估计分布函数为：

$$F(x)=\sum_{j=1}^{i}p_j \quad (3)$$

$$(1<x<=i \quad i=1,2,\ldots,t)$$

2.2 Hash 函数的构造

可按照如下步骤构造 Hash 函数：

① 首先对总体数据进行简单随机抽样，抽样针对各维变量进行。

② 按照(1)、(2)、(3)式得到各维变量的近似分布，构造 Hash 函数如下：

$$H(x_1,x_2,...,x_m)=F(x_1)F(x_2)...F(x_m) \quad (4)$$

(1)(2)(3)式实际上假定了各变量之间相互独立。由此可得出一个命题：

$x_1,x_2,....x_m$ 彼此之间相互独立时，$H(x_1,x_2,....x_m)$ 为变量 $X=X(x_1,x_2,....x_m)$ 的联合分布函数。该命题由独立随机变量的联合分布函数的性质即知。

3 Hash函数聚类的思想

假设数据流 S 上基于时间的滑动窗口 W 的时间跨度为 T，则可使用多个连续的滑动窗口，可简记为 $[t-T,t]$，$[t-T+\Delta,t+\Delta]$，…，$[t-T+(n-1)\Delta,t+(n-1)\Delta]$. 其中 Δ 的取值范围为 $[0,T]$，即每两个相邻的滑动窗口的时间跨度为 Δ，这样保证在对两个连续的滑动窗口中的数据采样时，有重叠的数据存在，以分析两个连续滑动窗口的数据流变化情况。

根据上述定义，要保证采样结果在允许的误差范围内，可以采用 Hash 函数对每个窗口内的数据进行采样。因为 $H(x_1,x_2,....x_m)$ 为变量的随机分布函数，所得的结果能近似于总体数据的随机分布。

当数据流的数据充满一个滑动窗口时，用 Hash 函数从中抽取样本数据量为 s，对这 s 个样本数据实施 PAM 算法。根据变量的三种类型，可定义最终聚类簇为 k 个（$k=3*m$，且 m 为整数），每种变量有 m 个聚类簇。在 $[t-T,t]$ 时刻，即在第一个滑动窗口内，对于整个数据样本中的每一个个体 X_j，判断 k 个中心点中哪一个与 X_j 最相似，得到聚类结果 $A^i=\{A_1^i,A_2^i,....A_k^i\}$。并将之保存在一个预先设定的存储池中。

同理，对随后的第二个滑动窗口也进行类似抽样。跟第一个窗口相比，由于其时间延迟为 Δ，所以相当于在时间 Δ 内，进入了一批数据，将最先进入第一个滑动窗口的时间内的数据给丢弃了，所以第二个窗口的采样结果与第一个窗口的采样结果相比，聚类簇中的数据发生了变化。假设第一个滑动窗口中抽样得到的数据有 a 个属性，而第二个滑动窗口中的数据有 b 个属性的数据在第一个中不存在，于是将这 b 个数据添加到存储池中，并保存到聚类结果 A^i 中。如此进行下去，直到 n 个时间窗口中抽样后的所有不同数据的属性全部保存下来。此时，高速、实时的流数据保存到存储池后变成静态数据，这些静态样本数据能较好的反映总体数据的属性。

4 数据流聚类算法HSCS

输入：从任意时刻 t 开始流入到滑动窗口内的数据。

输出：k 个聚类簇 $A^i=\{A_1^i,A_2^i,...,A_k^i\}$。

步骤：

① 根据 $F(x_1)$，$F(x_2)$，$F(x_3)$ 估计各数 据变量的分布函数。

② 构造 Hash 函数 $H(x_1,x_2,...,x_m)$。

③ 从任意时刻 t 开始，在时间跨度为 $[t-T,t]$ 的滑动窗口中，根据 Hash 函数对数据样本进行采样。

④ 把采样后的数据采用 PAM 算法将之分成 k 个聚类簇，并确定每个聚类簇的中心点。得到聚类结果 $A^i=\{A_1^i,A_2^i,...,A_k^i\}$，将数据保存在存储池中。

⑤ 对随后的滑动窗口用同样的方法采样，将每次的采样结果与之前的结果分析比较。

⑥ 数据流中的数据处理完毕之后，保存在存储池中的数据变为静态数据，对之采用常规的聚类算法进行聚类

5　实验与性能分析

本文所有的实验均在 PC 机上完成，其实验环境为：Intel Pentium IV，512M 内存，Windows XP SP2，采用 VC++6.0 实现。算法采用 KDD-CUP’99 真实数据集进行仿真。该数据集曾被多篇数据流聚类文章引用，它由一个局域网中 TCP 连接的原始记录所构成，各记录含有连接的持续时间、从源到目的的传输的字节数等。

本文从数据集中读取 15000 个数据，并设定最终聚类簇数量为 =30。分别计算出当样本数据量为 1000、3000、5000、8000、15000 时的算法运行时间，所得的实验结果如下图所示，从图中可以看出，当数据很小时，CluStream 算法运行时间较少，处理数据速度较快；但是当数据量逐渐增大，HSCS 算法的花费时间要少的多。

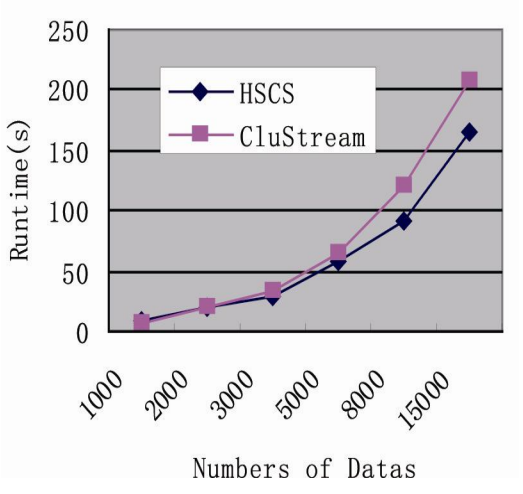

图 1　两种算法的运行时间比较

6　结论与进一步研究的方向

本文针对 CluStream 算法存在的一些不足，做了如下工作：

① 提出了 Hash 函数抽样的新思想。

② 提出了等时间跨度的滑动窗口集合的概念，针对不同时间下，不同滑动窗口中的数据流属性(或状态)变化情况，预测后来数据的属性(或状态)。

③ 实验证明，在处理含有大量数据的数据流时，HSCS 算法在处理时间上优于 CluStream 算法的。

目前的实验是针对数值型数据进行的，因此在下一步的工作中，我们将致力于多媒体数据、图形图像数据的聚类处理，以期获得更好的聚类算法和处理速度。

参考文献

1 Aggarwal CC, Han J. A Framework for Clustering Evolving Data Streams. Proc the 29th VLDB Conference. Berlin: Johann Christoph Freytag, 2003: 81－92.

2 Aggarwal CC , Han J. On High Dimensional Projected Clustering of Data Streams. Data Mining and Knowledge Discovery, 2005:252－273,

3 朱蔚恒,印鉴,谢益煌.基于数据流的任意形状聚类算法.软件学报, 2006,13(7):379－387.

4 张龙波,李战怀,等.带权值数据流滑动窗口随机抽样算法的改进.计算机工程与应用,2007,43(35):18－20.

5 常建龙,曹锋,周傲英.基于滑动窗口的进化数据流聚类. 软件学报, 2007,18(4):905－918.